# Homework 4

The purpose of this homework is to practice using randomization methods to run hypothesis tests. Please fill in the appropriate code and write answers to all questions in the answer sections, then submit a compiled pdf with your answers through gradescope by 11:30pm on Sunday September 29th.

## Part 1: Further exploration of OkCupid users' income

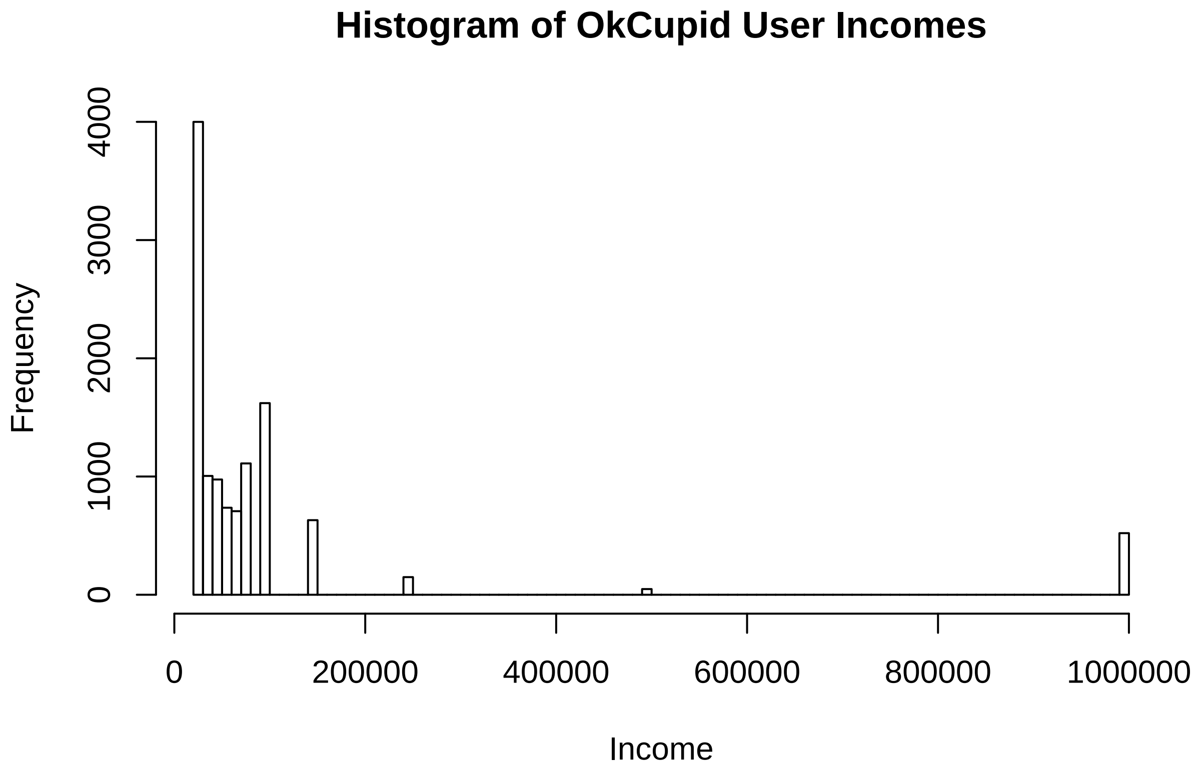
**Part 1.1 (5 points)** In homework 3 (problem 2.2) you calculated the population mean income for OkCupid $\mu$ assuming that the population consisted of only OkCupid users in the profiles data frame (and only those users who reported their income). The value for the population parameter for the mean income $\mu$ that you got should have been around $\$100,000$. The question then becomes, do we really believe that a typical OkCupid users is making around $\$100k$ a year?

Before diving into inferential statistics (such as creating confidence intervals and running hypothesis test) we really should always explore and visualize the data first. So let's do some exploratory analysis now, which will also be a useful review of some of the material we have already covered in this class.

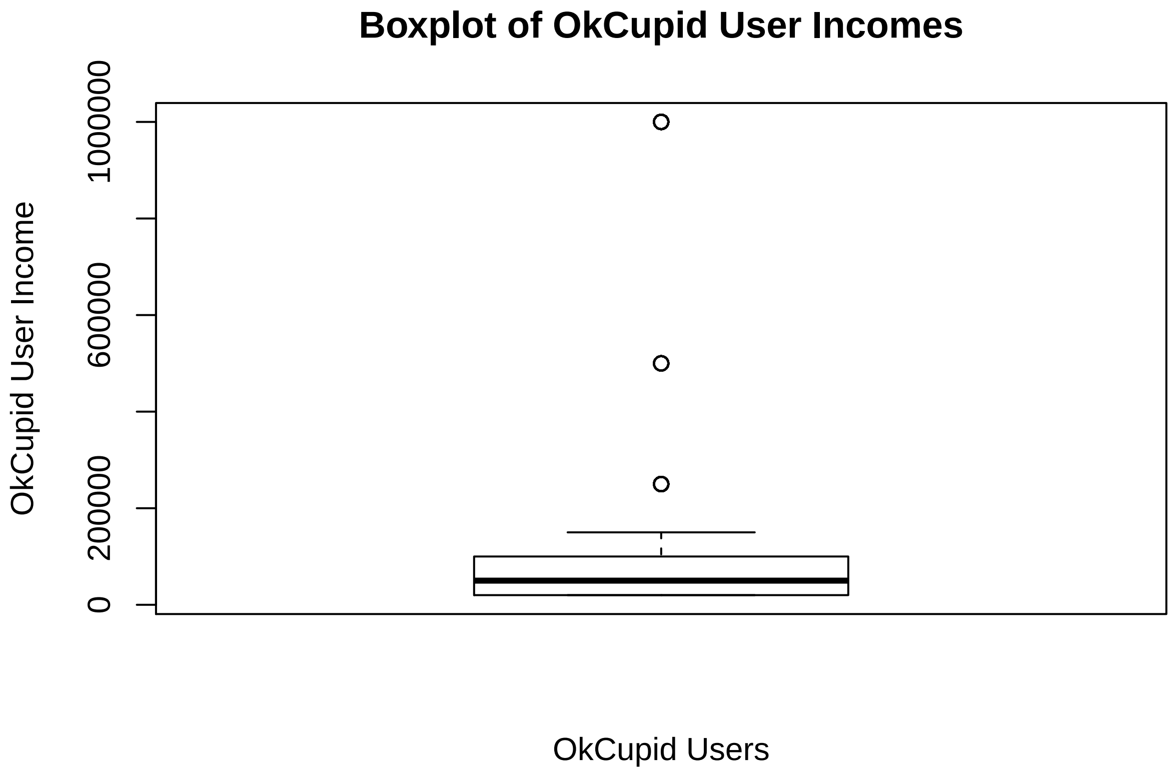
To start, create a histogram and a boxplot of the income data from the OkCupid users. Also, calculate the proportion of users who claimed their income to be a million dollars or more, and report the value below. Does the portion of people who say they are making a million dollars or more match what is reported in surveys of Americans? (use internet sources to get an estimate of how many Americans make a million dollars or more a year).

```r
library(okcupiddata)

all_income <- na.omit(profiles$income)
hist(all_income, main = "Histogram of OkCupid User Incomes",
     xlab = "Income", ylab = "Frequency", nclass = 100)
```



```r
boxplot(all_income,
xlab = "OkCupid Users",
ylab = "OkCupid User Income",
main = "Boxplot of OkCupid User Incomes")
```



```
#Proportion of million-dollar earners
(million_prop <- sum(all_income >= 1000000)/length(all_income))
```

```
## [1] 0.0452886
```

**Answer**

According to this data, 4.5% of OkCupid users are million-dollar earners. According to data from the university of Minnesota, in 2018 the 99th percentile of income in the United States was around $\$300,000$. This means less than 1% of people in the United States make more than 1 million a year. This is far below the proportion calculated from the reports of OkCupid users.

**Part 1.2 (5 points)** Now let's examine how the mean statistic is affected if we remove the users who reported making a milion dollars or more. To do this, compare the population mean income (i.e., mean of all valid values in the whole profiles data frame) when users who are making a million or more are excluded, to the mean income when all users are included, and create side-by-side boxplots of these two populations with the outliers removed from the plots. Report what these mean values are and whether the millionaires have a big impact on our estimates of the mean income. Also, report whether the boxplots look similar.

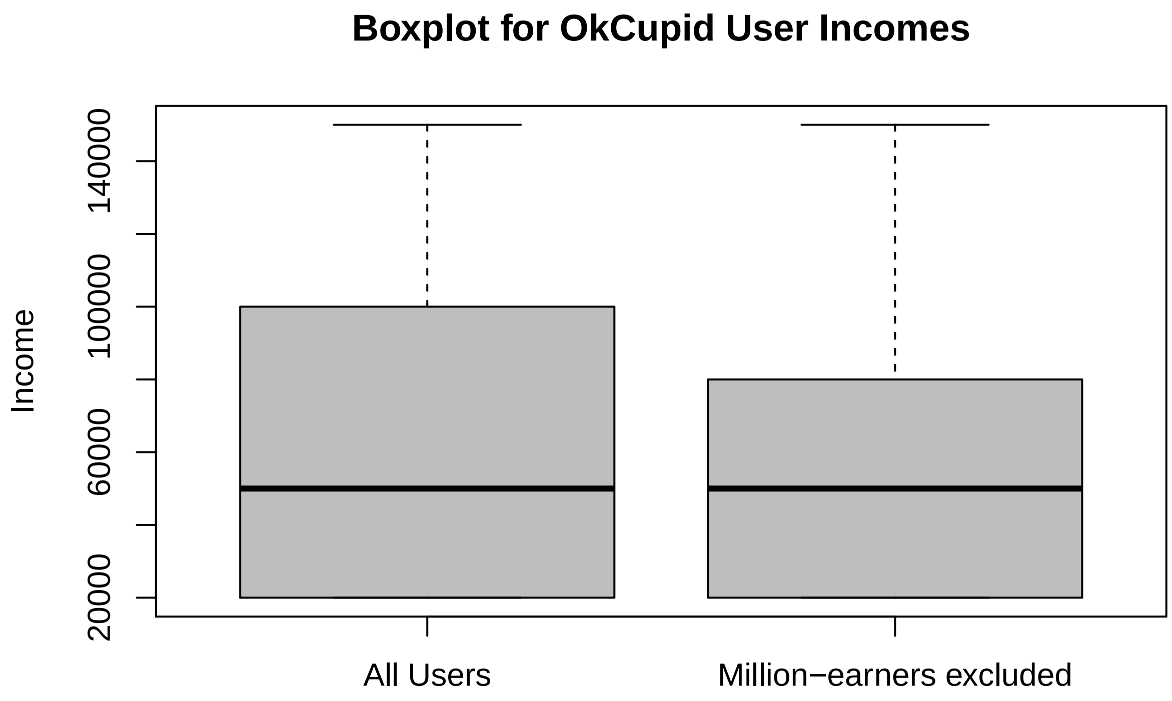
```r
#Mean for all users
all_income <- na.omit(profiles$income)
(mean(all_income))
```

```
## [1] 104395
```

```r
#Mean for users, million earners excluded
million_excl <- all_income[all_income < 1000000]
(mean(million_excl))
```

```
## [1] 61910.22
```

```r
#Comparative Boxplot
boxplot(all_income, million_excl, outline = FALSE, col = "gray",
        main = "Boxplot for OkCupid User Incomes",
        names = c("All Users", "Million-earners excluded"),
        ylab = "Income")
```



**Answers**

The mean income for all users was $104,385, wheras the mean income for users excluding millionares is $61,910.22. The presence of millionares on the data set does have a large impact on the mean income (the mean was increased about $40,000). However, when we look at the boxplots it seems that they are very similar. The medians are very close, likely because the median is robust to outliers (the removal of the million-dollar values most likely did not change the middle value of the set). There is some difference in that the third quartile is smaller when millionaires are excluded, which is probably because the third quartile is likely to be affected by the removal of the highest values.

**Part 1.3 (10 points)** As we saw above, the mean statistic is not very resistant to outliers (i.e. it can be heavily influenced by large values). This can lead to estimates that are not really representative of what we might think of as a "typical American". The median statistic is resistant to large values however, and so might be more meaningful here.

Let's examine the median by doing the following:

1) Calculating the median income on the whole okcupid data set a) including and b) excluding people who report making over one million dollars.

2) Use the bootstrap percentile method to create a confidence interval for the median using the first 50 OkCupid users. Again, do this a) including and b) excluding people who report making over one million dollars. (Note: confidence intervals are always computed on a sample of data (here of size n = 50), and the purpose of this exercise is to see whether our confidence interval based on using the bootstrap percentile method captures the true population median parameter calculated from all OkCupid users).

For the confidence interval, you only need to show your code when including all users (part a's) but fill in the table below reporting the values. Alternatively, create a function for calculating bootstrap confidence intervals and run your code twice with different the million dollar incomes included vs. excluded.

Fill in the table below showing the mean values calculated above from the whole population, as well as the median values and the median confidence intervals based on 50 users, and describe whether the results for the median change much depending on whether the millionares' incomes are excluded. Also describe whether the median seems like a better description of a "typical income" compared to the mean, and if there might be any issues trusting the confidence intervals that were created.

```r
#Median of entire dataset
median(all_income)
```

```
## [1] 50000
```

```r
#Median of data with exclusion
median(million_excl)
```

```
## [1] 50000
```

```r
#Bootstrap percentile function
bts_per_func <- function(income_vec){
bootstrap_dist_income <- c()
for (i in 1:10000){
  bootstrap_dist_income[i] <- median(sample(income_vec[1:50], replace = T))
}
(CI_boot_per <- quantile(bootstrap_dist_income, c(.025, .975)))
}

bts_per_func(all_income)
```

```
##  2.5% 97.5%
## 50000 80000
```

```r
bts_per_func(million_excl)
```

```
##  2.5% 97.5% 
## 50000 80000 
```

**Answer**

| | pop mean | pop median | CI median |
|---|---|---|---|
| all data | 104395 | 50000 | (50000, 80000) |
| exluding millionaire income | 61910.22 | 50000 | (50000, 80000) |

The population median as well as Confidence Interval do not not change whether or not we remove the millionares in the data set. In this way it seems to be a better descriptor of income compared to the mean, since the median captures the "middle" value of the income without being influenced by the outliers, as the mean statistic would. However, there may be some issues trusting these confidence intervals because we notice that the population median both including and excluding millionaires is only at the end of the interval. Also, taking a sample of only 50 users causes there to be a wider interval since the sample may not accurately represent the entire population.

## Problem 2: Hypothesis tests and confidence intervals for a single proportion

Paul the Octopus was an octopus who became famous for predicting winners of soccer matches during the 2010 World Cup. To examine Paul's psychic abilities, two containers of food (mussels) were lowered into the Paul's tank prior to each soccer game. The containers were identical, except for country flags of the opposing teams, one on each container. Whichever container Paul opened first was deemed his predicted winner.

Paul (in a German aquarium) became famous for correctly predicting 11 out of 13 soccer games during the 2010 World Cup. Let's use hypothesis testing to examine whether Paul is actually psychic or if he was merely guessing.

**Part 2.1 (5 points)**: State the null and alternative hypotheses testing whether Paul is psychic using both words and in the appropriate symbols. Also describe what the significance level means and denote it with the commonly used symbol and commonly used value.

**Answer**:

$H_0$ : Paul does not have the ability to predict the winners (His guessing ability is $\pi = 0.5$).

$H_A$ : Paul has the ability to predict the winners of soccer matches (His guessing ability is $\pi > 0.5$)

The significance level $\alpha$, which is commonly $\alpha = 0.05$, is the probability that we reject the null hypothesis when it is actually true. This means there is a 5% risk of concluding Paul is psychic when he is actually not.

**Part 2.2 (5 points)** : Compute the statistic of interest and save it in an object paul_stat. Do you think it is likely you would get a statistic this extreme if Paul was guessing?

```
# calculate the observed statistic
(paul_stat <- 11/13)
```

```
## [1] 0.8461538
```

**Answer**: If Paul was guessing, his success rate should be closer to 0.5, so it seems unlikely that he would have a success rate of $\hat{p} = 0.846$.

**Part 2.3 (5 points)** : Now use the rbinom() function to generate a null distribution that would occur if Paul was guessing, and save the results in an object called null_distribution.

Remember that the arguments to the rbinom(num_sims, size, prob) are:

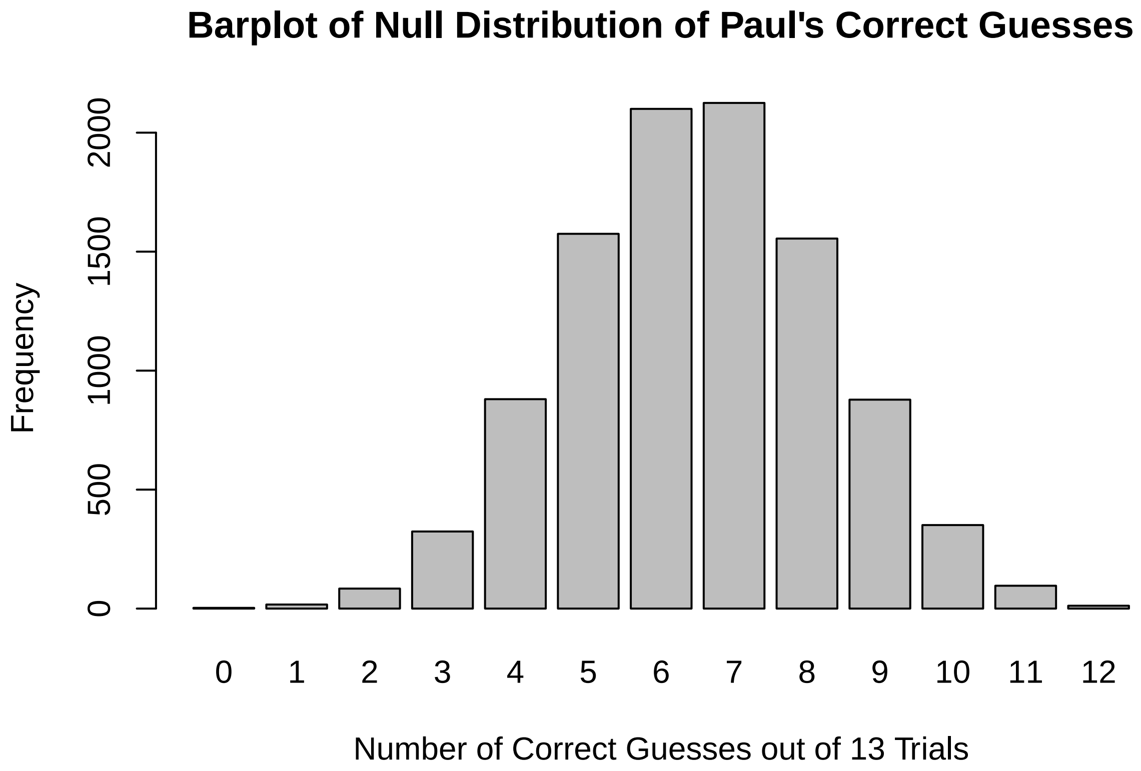
- num_sims: the number of simulations to run
- size: the number of "coin flips" in each simulation
- prob: the probability of getting heads on each coin flip.

Also create a table showing the number of "heads" each simulation produced, and plot this null distribution table as a bar plot.

```
null_distribution <- rbinom(10000, 13, 0.5)
table(null_distribution)
```

```
## null_distribution
##    0    1    2    3    4    5    6    7    8    9   10   11   12 
##    3   17   84  324  880 1575 2100 2125 1555  878  351   96   12
```

```
barplot(table(null_distribution),
        main = "Barplot of Null Distribution of Paul's Correct Guesses",
        xlab = "Number of Correct Guesses out of 13 Trials", ylab = "Frequency")
```



**Part 2.4 (5 points)**: Now use the variables null_distribution and paul_stat to calculate the number of simulations that had as many or more "heads" than as Paul's correct soccer prediction answers. Convert this to a p-value by dividing by the total number of simulations. Does this p-value provide evidence that Paul is psychic?

```
sum(null_distribution/13 >= paul_stat)
```

```
## [1] 108
```

```
#p-value
(p_value <- sum(null_distribution/13 >= paul_stat)/10000)
```

```
## [1] 0.0108
```

**Answer**: p-value = 0.0108. This is less than $\alpha = 0.05$, therefore the results are statistically significant and we are provided evidence that Paul is psychic.

**Part 2.5 (2.5 points)** Make a judgement call as to whether you believe Paul is psychic based on the p-value and any other information you think is releveant. Make sure to justify your answer to explain Paul's prediction abilities.

**Answer**:

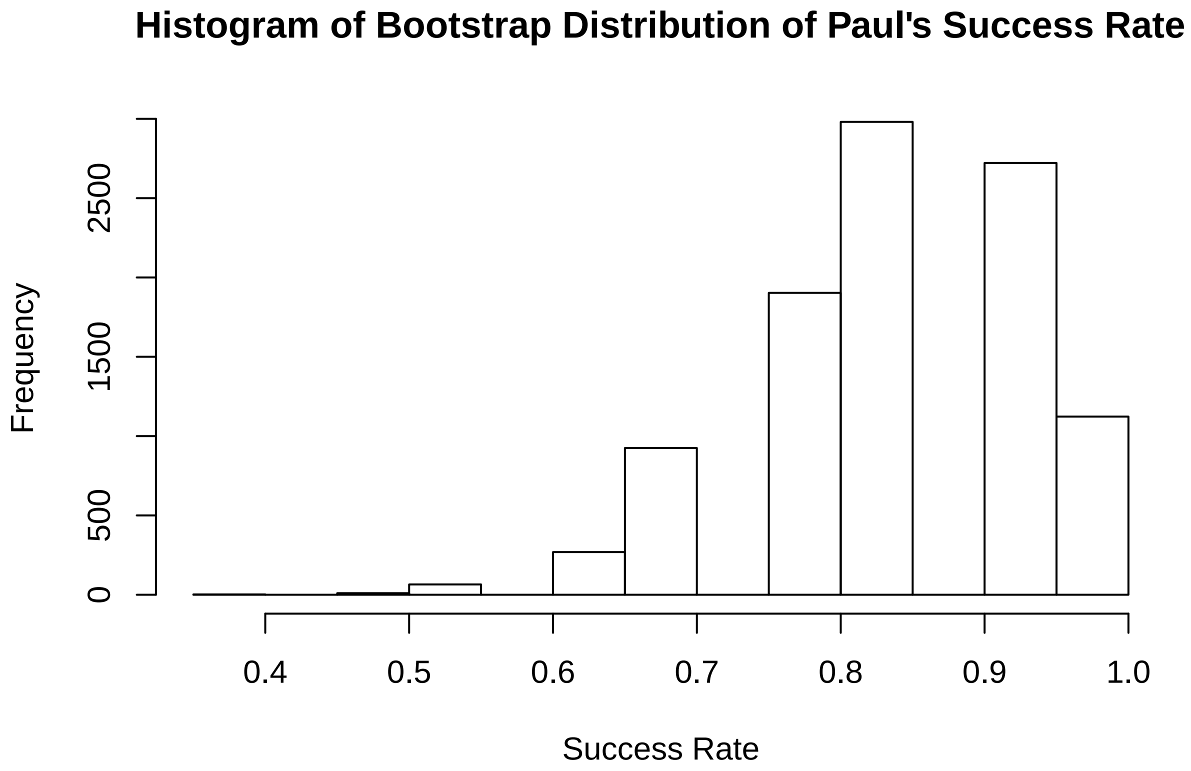
Our calculated p-value is less than the significance level. According to our p-value there is a 0.0108 probability that Paul could guess correctly more than 11 times, so theoretically we should reject the null hypothesis in favor of the alternative. However, with everything we know about modern day science and how the brain works, I highly doubt Paul is psychic. There are many factors that could have influenced Paul's decision, such as coloring or placement of the flags. A result like this would warrant further investigation until we can conclude that Paul is psychic.

**Part 2.6 (10 points)** Calculate the confidence interval for Paul's prediction abilities using the bootstrap methods

```
# a vector of Paul's answers whether they were correct or incorrect (not necessarily in the order)
pauls_answers <- c(rep('correct', 11), 'incorrect', 'incorrect')   # 11 correct, 2 incorrect


# continue creating the bootstrap distribution from here
bootstrap_dist_paul = c()
for (i in 1:10000){
  bootstrap_sample <- sample(pauls_answers, replace = TRUE)
  bootstrap_dist_paul[i] <- sum(bootstrap_sample == "correct")/length(bootstrap_sample)
}

# plot the boostrap distribution
hist(bootstrap_dist_paul,
     main = "Histogram of Bootstrap Distribution of Paul's Success Rate",
     xlab = "Success Rate", ylab = "Frequency")
```



```
# CI using the bootstrap based on a normal approximation to the bootstrap distribution
(CI_paul <- paul_stat + sd(bootstrap_dist_paul)*c(-2,2))
```

```
## [1] 0.6479314 1.0443762
```

```
# CI using the bootstrap percentile method
(CI_perc <- quantile(bootstrap_dist_paul, c(0.025, 0.975)))
```

```
##      2.5%     97.5%
## 0.6153846 1.0000000
```

**Answer**:

CI using the bootstrap distribution method: [0.6479314, 1.0443762]

CI using the bootstrap percentile method: [0.6153846, 1]

**Part 2.7 (5 points)** There is also a formula for calculating the standard error of a proportion which is:

$s_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Use this formula to create 95% confidence intervals and compare it to the 95% confidence intervals you calculated in the previous problem. Do they appear to be similar?

```r
se_formula <- sqrt(((paul_stat)*(1 - paul_stat))/13)
(CI_formula <- paul_stat + se_formula*c(-2,2))
```

```
## [1] 0.6460173 1.0462903
```

**Answer**

The confidence interval calculated from the formula is [0.6460173, 1.0462903]. This appears to be most similar to the one calculated using the bootstrap distribution method, although all three are pretty similar. Only the bootstrap percentile method does not exceed 1 (indicating a 100% success rate) on the upper end.

## Problem 3: Permutation tests for comparing two means - Sleep or Caffeine for Memory?

The consumption of caffeine to benefit alertness is a common activity practiced by 90% of adults in North America. Often caffeine is used in order to replace the need for sleep. One recent study compared students' ability to recall memorized information after either the consumption of caffeine or a brief sleep (see Mednick et al., 2018

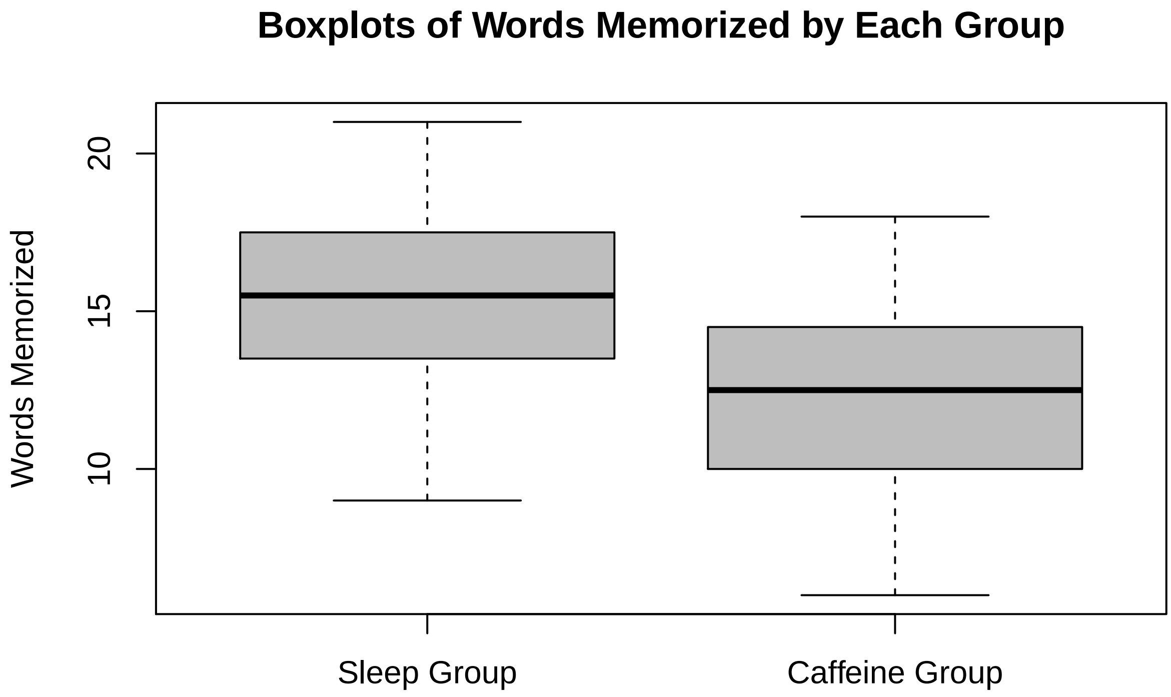
A random sample of 35 adults (between the ages of 18 and 39) were randomly divided into three groups and verbally given a list of 24 words to memorize. During a break, one of the groups took a nap for an hour and a half, another group was kept awake and then given a caffeine pill an hour prior to the testing, and a third group was given a placebo. The response variable of interest is the number of words participants are able to recall following the break.

Let's run a hypothesis test to see if there is statistically significant difference in the mean number of words recalled between the group that got *sleep* and the group that got *caffeine* (we will ignore the placebo group since I don't have data from that group).

**Part 3.0 (5 points)** The data for the number of words recalled by members in each of the two groups is below. Start by creating a side by side boxplot comparing the two groups. Describe below whether there appears to be a difference between the groups.

```r
sleep_condition <- c(14, 18, 11, 13, 18, 17, 21, 9, 16, 17, 14, 15)
caffeine_condition <- c(12, 12, 14, 13, 6, 18, 14, 16, 10, 7, 15, 10)

boxplot(sleep_condition, caffeine_condition,
        main = "Boxplots of Words Memorized by Each Group",
        names = c("Sleep Group", "Caffeine Group"), ylab = "Words Memorized", col = "gray")
```



**Answer**:

From the boxplot it seems that the group that slept was able to memorize more words than the group that took caffeine. Each quartile statistic is greater than that of the caffeine group.

In parts 3.1 to 3.5 you will now do the 5 steps to run a hypothesis test.

**Part 3.1 (5 points)** State the null and alternative hypotheses using words and symbols. Also describe the significance level is and denote it with the appropriate symbol.

**Answer**:

$H_0$ : There is no difference in the memorization ability of people who sleep versus caffeine.

$\mu_{sleep} - \mu_{caffeine} = 0$.

$H_A$ : There is a difference in the memorization ability of people who sleep versus caffeine

$\mu_{sleep} - \mu_{caffeine} \neq 0$.

Significance level: $\alpha = 0.05$

**Part 3.2 (5 points)** Calculate the value of that statistic for the observed sample, and use the appropriate symbol notation along with its value below.

```r
#Sleep group
mean(sleep_condition)
```

```
## [1] 15.25
```

```r
#Caffeine group
mean(caffeine_condition)
```

```
## [1] 12.25
```

```r
(obs_stat <- mean(sleep_condition) - mean(caffeine_condition))
```

```
## [1] 3
```

**Answer**:

$\bar{x}_{sleep} = 15.25$

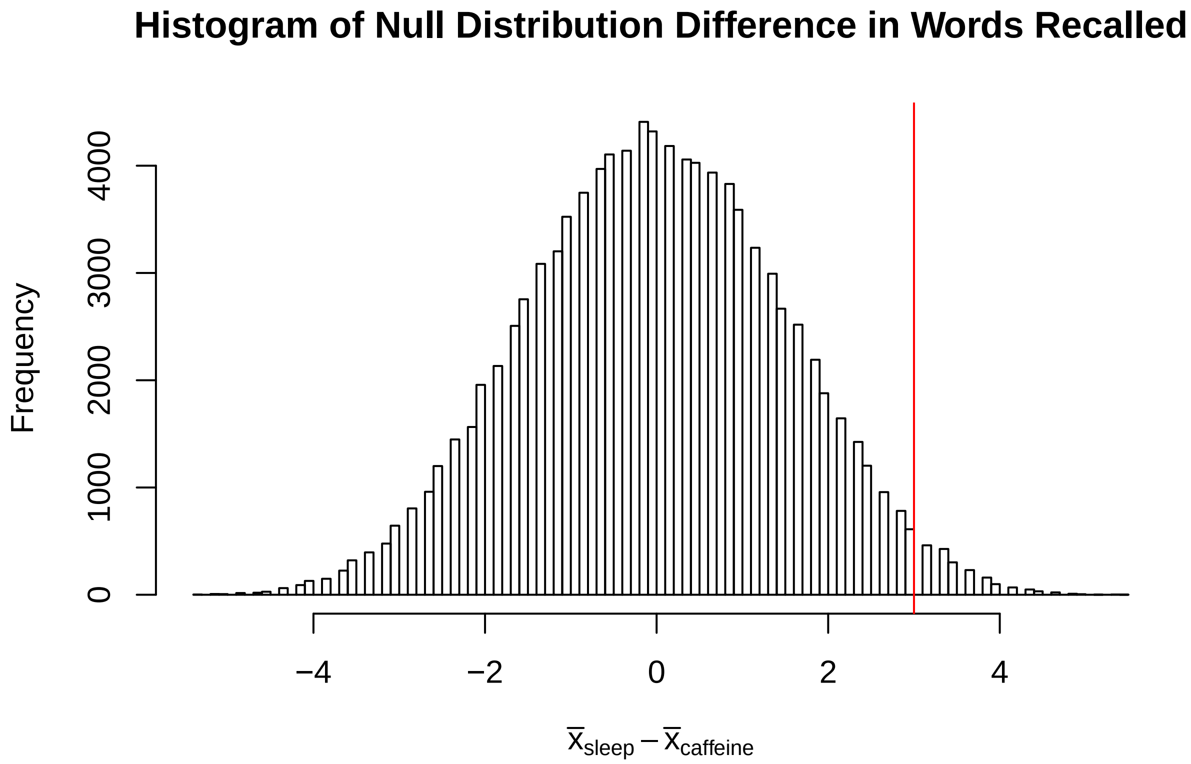
$\bar{x}_{caffeine} = 12.25$

$\bar{x}_{sleep} - \bar{x}_{caffeine} = 3$

**Part 3.3 (10 points)** Create a null distribution using a permutation test (i.e., combine data from both groups, randomly assign them to a fake "caffeine" and "sleep group", calculate a null statistic, and repeat 100,000 times to get a null distribution). Also plot a histogram of the null distribution and add a red vertical line to the plot at the value of the observed statistic.

```r
combined <- c(sleep_condition, caffeine_condition)
null_dist <- c()
for (i in 1:100000){
  shuff_data <- sample(combined)
  shuff_sleep <- shuff_data[1:12]
  shuff_caffeine <- shuff_data[13:24]

  null_dist[i] <- mean(shuff_sleep) - mean(shuff_caffeine)
}
hist(null_dist, main = "Histogram of Null Distribution Difference in Words Recalled",
     xlab = TeX(("$\\bar{x}_{sleep} - \\bar{x}_{caffeine}$")),
     ylab = "Frequency",
     nclass = 100)
abline(v = obs_stat, col = "red")
```



**Part 3.4 (5 points)** Now calculate the p-value in the R chunk below.

```
(p_val <- sum(null_dist >= obs_stat, null_dist <= -obs_stat)/100000)
```

```
## [1] 0.05049
```

**Part 3.5 (2.5 points)** Are the results statistically significant? What would we conclude if we used a strictly "Neyman-Pearson paradigm" where we only reject results that are less than our significance level? Do you believe there is a difference between these groups?

**Answers**:

The p-value is 0.05123, which is greater than our significance level $\alpha = 0.05$, so under the Neyman-Pearson paradigm the we cannot concluded that the results are statistically significant and we fail to reject the null hypothesis that there is no difference between the groups. However, I do believe there is a difference between the groups based on past scientific evidence which consistently confirms the importance of sleep for memory retention, as well as the fact that the p-value is very close to $\alpha$. The boxplots do show a difference in the values as well. If this test was a one-tailed test attempting to find whether $\bar{x}_{sleep} > \bar{x}_{caffeine}$, we would find that the p-value is smaller than $\alpha$.

**Part 3.6 (5 points)** Parametric hypothesis tests are hypothesis tests where the null distribution is given by a mathematical density function. When comparing two means, a parametric hypothesis test, that you likely learned about in introductory statistics, is the t-test, where the null distribution is a t-distribution.

R has a built in function called `t.test(sample1, sample2)` that takes two samples of data and runs a t-test on them. Use this function to compare the sleep and caffeine groups and report if there is a statistically significant difference between the groups. Also report the 95% confidence interval that the t.test function returns and describe whether this confidence interval is consistent with it being plausiable that there is no difference between the population means of these two groups.

```r
t.test(sleep_condition, caffeine_condition)
```

```
##
##  Welch Two Sample t-test
##
## data:  sleep_condition and caffeine_condition
## t = 2.1438, df = 21.894, p-value = 0.04342
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.09699633 5.90300367
## sample estimates:
## mean of x mean of y
##     15.25     12.25
```

**Answers**:

The p-value is 0.04342, which is less than $\alpha = 0.05$, so we can conclude that the results are statistically significant and there is evidence that there is a difference between the groups. The 95% confidence interval is [0.097, 5.903]. Since this interval only includes positive values and no 0 value, it not consistent with it being plausible that there is no different in the means of the two groups. Instead, this interval is consistent with the hypothesis that words memorized by the sleep group are great than the number of words memorized by the caffeine group.

## Reflection (5 points)

Please reflect on how the homework went by going to Canvas, going to the Quizzes link, and clicking on Reflection on homework 4